A Phase II Clinical Trial with Efficacy and Toxicity Outcomes and Baseline Covariates

Brock, Kristian; Billingham, Lucinda; Yap, Christina; Middleton, Gary

A Phase II Clinical Trial Design for Associated Co-Primary Efficacy and Toxicity Outcomes with Baseline Covariates

Kristian Brock, Lucinda Billingham, Christina Yap and Gary Middleton

Abstract The experimental design presented here is motivated by a phase II clinical trial called PePS2, investigating the efficacy and safety of an immunotherapy called pembrolizumab in a specific subgroup of lung cancer patients. Previous trials have shown that the probability of efficacy is correlated with particular patient variables. There are clinical trial designs that investigate co-primary efficacy and toxicity outcomes in phase II, but few that incorporate covariates. We present here the approach we developed for PePS2, latterly recognised to be a special case of a more general method originally presented by Thall, Nguyen and Estey. Their method incorporates covariates to conduct a dose-finding study but has been scarcely used in trials. Dose-finding is not required in PePS2 because a candidate dose has been widely tested. Starting from the most general case, we introduce our method as a novel refinement appropriate for use in phase II, and evaluate it using a simulation study. Our method shares information across patient cohorts. Simulations show it is more efficient than analysing the cohorts separately. Using the design in PePS2 with 60 patients to test the treatment in six cohorts determined by our baseline covariates, we can expect error rates typical of those used in phase II trials. However, we demonstrate that care must be taken when specifying the models for efficacy and toxicity because more complex models require greater sample sizes for acceptable simulated performance.

Key words: covariate, efficacy, phase II, toxicity, trial

Kristian Brock
University of Birmingham, UK, B15 2TT, e-mail: k.brock@bham.ac.uk

Lucinda Billingham
University of Birmingham, UK, B15 2TT, e-mail: l.j.billingham@bham.ac.uk

Christina Yap
University of Birmingham, UK, B15 2TT, e-mail: c.yap@bham.ac.uk

Gary Middleton
University of Birmingham, UK, B15 2TT, e-mail: g.middleton@bham.ac.uk

1 Introduction

There is a relative dearth of phase II clinical trial designs that incorporate patient covariates to assess efficacy and toxicity. We introduce a novel approach here.

Our motivation is a phase II trial called PePS2 that investigates an immunotherapy in a specific subgroup of lung cancer patients. We developed a Bayesian regression method that adjusts for predictive patient data available at trial commencement to investigate co-primary binary outcomes. We latterly learned that our design is a special case of Thall, Nguyen & Estey (TNE), a family of methods that perform dose-finding trials guided by efficacy and toxicity outcomes whilst accounting for baseline patient data [17] . Their design yields personalised dose recommendations.

PePS2 is not a dose-finding trial. Instead, it seeks to estimate the probabilities of efficacy and toxicity at a dose of pembrolizumab previously demonstrated to be safe and effective in a closely-related group of patients [9] . To acknowledge its heritage, we introduce our design as a novel simplification of TNE that removes the dose-finding components so that it may be used in phase II.

In Section 2, we describe the PePS2 trial and the pertinent clinical data from previous trials. In Section 3, we review the literature for suitable experimental designs. We describe our design in detail in Section 4 and evaluate it with a simulation study in Section 5. Finally, in Section 6, we describe future plans for this work.

2 The Clinical Trial Scenario

PePS2 is a phase II trial of pembrolizumab in non-small-cell lung cancer (NSCLC) patients with Eastern Cooperative Oncology Group performance status 2 (PS2). NSCLC is a common sub-type of lung cancer. Patients with PS2 are ambulatory and capable of self-care but typically too ill to work. Critically, it is doubtful that a PS2 patient could tolerate the toxic side effects of chemotherapy.

The primary objective of the trial is to learn if pembrolizumab is associated with sufficient disease control and tolerability to justify use in PS2 patients. The joint primary outcomes are (i) *toxicity*, defined as the occurrence of a treatment-related dose delay or treatment discontinuation due to adverse event related to pembrolizumab; and (ii) *efficacy*, defined as the occurrence of stable disease, partial response (PR) or complete response (CR), without prior progressive disease, at or after the second post-baseline disease assessment by version 1.1 of the *Response Evaluation Criteria In Solid Tumors* [8] . The second assessment is scheduled to occur at week 18.

Pembrolizumab inhibits the programmed cell death 1 (PD-1) receptor via the programmed death-ligand 1 (PD-L1) protein. It has been shown to be active and tolerable in patients with better performance status [9] . Overall, 19.4% of patients had an objective response (PR or CR) and 9.5% experienced a major adverse event, defined as an event of at least grade 3 by the *Common Terminology Criteria for Adverse Events*, v4.0. These statistics compare favourably to those typically seen in

advanced NSCLC patients under chemotherapy [1, 13] . We foresee no reason why they should be materially different in PS2 patients.

Table 1 Objective response rates for the validation sample ($n = 204$) in [9] .

PD-L1 Group	Criteria	Objective Response %, (95% CI)
Low	PD-L1 score $< 1\%$	10.7 (2.3, 28.2)
Medium	$1\% \geq$ PD-L1 score $< 50\%$	16.5 (9.9, 25.1)
High	PD-L1 score $\geq 50\%$	45.2 (33.5, 57.3)

Garon *et al.* introduce the PD-L1 proportion score biomarker, defined as the percentage of neoplastic cells with staining for membranous PD-L1[9] . Efficacy outcomes for the 204 patients in their validation group, summarised by PD-L1 score, are shown in Table 1. Objective responses are observed in all cohorts and the rate increases with PD-L1. Based on this information, we expect PD-L1 to be predictive of response in our PS2 population.

Furthermore, 24.8% of patients who had received no previous anti-cancer therapy (treatment-naive, TN) achieved a response, compared to 18.0% in the group that had been previously treated (PT) [9] . This represents a potentially small but important effect that should be considered when testing the treatment. We propose to investigate pembrolizumab by jointly stratifying by the three Garon PD-L1 groups, and PT and TN statuses. Each patient will belong to exactly one of six cohorts, as demonstrated in Table 2.

Cohort	Previous treatment status	PD-L1 category	$x_i = (x_{1i}, x_{2i}, x_{3i})$
1	TN	Low	(0,1,0)
2	TN	Medium	(0,0,1)
3	TN	High	(0,0,0)
4	PT	Low	(1,1,0)
5	PT	Medium	(1,0,1)
6	PT	High	(1,0,0)

Table 2 Cohorts used in the PePS2 trial. x_i shows the predictive variable vector for patient i.

In phase II, there is strong motivation to deliver findings quickly to inform the next study phase. Recruitment of approximately 60 PS2 patients within one year would be feasible but accrual materially higher would be unlikely. Given the relative dearth of treatment alternatives, we seek to offer the trial to all PS2 patients and not stratify accrual. Pembrolizumab has not been investigated in PS2 patients so the clinical scenario requires a trial design that tests efficacy and toxicity. Given the evidence that PD-L1 and pretreatedness are associated with response, it is highly desirable to use a trial design that incorporates this predictive information. The next section describes our search for a clinical trial design to achieve these objectives.

3 Review of Available Trial Designs

We sought a clinical trial design that uses covariates to study co-primary binary outcomes. The well-known phase II design by Bryant and Day (BD) takes threshold rates of efficacy and toxicity and returns the number of events to approve the treatment [4] . For given levels of significance and power, the thresholds identify the optimal trial of the competing outcomes. The design does not use covariates, assuming the population to be homogeneous. Parallel BD designs in our six cohorts would require a prohibitively large total sample size. Other phase II sequential designs with multiple outcomes [3, 6, 7, 11, 14, 15] generally focus on providing stopping rules rather than incorporating predictive information.

Several phase I dose-finding designs [2, 16, 19] use co-primary outcomes. These could potentially be adapted to our purpose, although they generally do not use covariates. A notable exception is TNE, an extension of EffTox [16] that adds patient covariates to analyse co-primary efficacy and toxicity at different doses. The objective of their Bayesian design is to recommend a personal dose of an experimental agent, after adjusting for baseline data. The design was used in a dose-finding study of PR104 in relapsed or refractory acute myeloid or lymphoblastic leukaemia [12] . We found no other examples of its use, and no suggestion that it had been adapted for the non-dose-finding context. Our proposed design can be considered as a simplification of TNE for use in phase II.

4 Assessing Efficacy and Toxicity and Adjusting for Covariates

In this section, we describe the statistical design used in PePS2, with the general TNE model as the starting point. We call this design P2TNE, for *Phase II Thall, Nguyen & Estey*. TNE present marginal probability models for an experimental treatment:

$$\text{logit}\,\pi_k(\tau, x, y, \theta) = f_k(\tau, \alpha_k) + \beta_k x + \tau\gamma_k y\,, \tag{1}$$

where $k = E, T$ denote efficacy and toxicity respectively. τ is the given dose appropriately normalised; x and y are vectors of covariates, with y interacting with dose; θ is a pooled vector of all parameters to be estimated; $f_k(\tau, \alpha_k)$ characterise the dose effects; and β_k and γ_k are vectors of covariate effects and dose-covariate interactions. TNE also introduce similar models for the events under historical treatments by which informative data on dose and covariate effects can be incorporated.

The authors consider joint models for associating events. They present an example using the Gumbel model, as used in [16] :

$$\begin{aligned}\pi_{a,b}(\pi_E, \pi_T, \psi) = (\pi_E)^a(1-\pi_E)^{1-a}(\pi_T)^b(1-\pi_T)^{1-b} \\ + (-1)^{a+b}(\pi_E)(1-\pi_E)(\pi_T)(1-\pi_T)\frac{e^\psi - 1}{e^\psi + 1}\,,\end{aligned} \tag{2}$$

where a and b equal 1 when efficacy and toxicity occur in a given patient respectively, else 0. For $\psi \in \mathbb{R}$, the fractional term takes values on (-1, 1), reflecting the correlation between the events. We refer to ψ as the association parameter.

To derive P2TNE, we remove all terms related to τ in (1) to reflect that dose is fixed. Furthermore in PePS2, we consider only the historic outcomes of the same single experimental treatment under a closely-related cohort of patients with NSCLC.

Let $x_i = (x_{1i}, x_{2i}, x_{3i})$ denote the covariate data and a_i, b_i the occurrence of efficacy and toxicity in patient i. For trial data:

$$X = \{(x_1, a_1, b_1), ..., (x_n, a_n, b_n)\} ,$$

the aggregate likelihood function is

$$\mathscr{L}(X, \theta) = \prod_{i=1}^{n} \pi_{a_i,b_i}(\pi_E(x_i, \theta), \pi_T(x_i, \theta), \psi) .$$

Let θ have prior distribution $f(\theta)$. For patients with covariate data x, the posterior expectation of the probability of efficacy under treatment is

$$\mathbb{E}(\pi_E(x, \theta)|X) = \frac{\int \pi_E(x, \theta) f(\theta) \mathscr{L}(X, \theta) d\theta}{\int f(\theta) \mathscr{L}(X, \theta) d\theta} ,$$

and the posterior probability that the rate of efficacy exceeds some threshold π_E^* is

$$\Pr(\pi_E(x, \theta) > \pi_E^*|X) = \frac{\int \mathbb{I}(\pi_E(x, \theta) > \pi_E^*) f(\theta) \mathscr{L}(X, \theta) d\theta}{\int f(\theta) \mathscr{L}(X, \theta) d\theta} .$$

The treatment is acceptable in patients with covariate vector x if

$$\begin{aligned} \Pr(\pi_E(x, \theta) > \pi_E^*|X) &> p_E \\ \Pr(\pi_T(x, \theta) < \pi_T^*|X) &> p_T , \end{aligned} \tag{3}$$

where π_E^*, p_E, π_T^* and p_T are chosen by the trialists. The clinical investigator chose the values $\pi_E^* = 0.1$ and $\pi_T^* = 0.3$ to reflect that efficacy less than 10% or toxicity exceeding 30% would render the treatment unattractive for further study in this patient group. We derived $p_E = 0.7$ and $p_T = 0.9$ by simulation using the method described below. Our chosen models for marginal efficacy and toxicity are:

$$\begin{aligned} \operatorname{logit} \pi_E(x_i, \theta) &= \alpha + \beta x_{1i} + \gamma x_{2i} + \zeta x_{3i} \\ \operatorname{logit} \pi_T(x_i, \theta) &= \lambda , \end{aligned} \tag{4}$$

with the events associated by (2). Our efficacy model assumes that the event log-odds for PT patients in the PD-L1 categories are a common linear shift of those in TN patients, an assumption we call *piecewise parallelism*, broadly supported by [9]. The rate of toxicity is assumed uniform across groups, supported by the data in [9, 10] . We analyse more complex models that relax each of these assumptions.

5 Simulation Study

Choice of priors is contentious in clinical trials. We simulated performance under diffuse, regularising, and informative priors. Our diffuse priors are normal with $\mu = 0$ and $\sigma = 10$. Regularising priors expect event rates close to 20% in all cohorts, put the majority of prior predictive mass in the left tail, but admit that event rates can be high. Informative priors expect event rates similar to those observed in [9], modestly penalised to reflect PS2 patient prognosis.

Table 3 shows operating characteristics using 60 patients. We tuned p_E and p_T by simulation in key benchmark scenarios, requiring that the design approve in all cohorts: (i) with at least 80% probability in scenario 1; and (ii) with no more than 5% probability in scenario 2. These probabilities reflect typical values for frequentist power and significance in phase II trials. Starting with $p_T = p_T = 0.7$, we saw that the designs accepted too often in scenario 2. With patients potentially near end-of-life, we chose to adjust operating performance by increasing certainty when evaluating toxicity; $p_T = 0.8$ was still too permissive but $p_T = 0.9$ achieved our goal under the regularising and diffuse priors, and $p_T = 0.95$ under informative priors. Scenarios 4-6 show that performance is good in settings inspired by the reported data [9, 10] . Compared to diffuse priors, the regularising priors improve approval probability without pre-empting covariate effects like the informative priors.

Table 3 also shows performance of beta-binomial conjugate models applied to cohorts individually with $Beta(1,1)$ priors, accepting if (3) is satisfied with $p_E = 0.7$ & $p_T = 0.9$. By incorporating baseline covariates, P2TNE considerably improves performance without erroneously inflating acceptance in scenarios 2 and 5.

The diffuse priors generate prior predictive distributions with most of the probability mass polarised close to events rates of 0 and 1, inconsistent with our beliefs and the published data. Coverage of posterior credible intervals was lowest and empirical standard error of estimates highest under the diffuse priors (data not shown).

Our model choices (4) imply fairly strong assumptions. We analyse model embellishments to infer the cost of greater model complexity. We relax the piecewise parallel assumption by adding interactions terms to the efficacy model. Under diffuse priors, approval probabilities and coverages decrease in our scenarios. An extra 20-40 patients are required to match performance of the simpler model under diffuse priors. To correctly improve the rejection probability in cohort 4 under scenarios 4-6, this model requires several times the initial sample size, an unjustifiable increase.

We relaxed the assumption that toxicity is uniform over groups by mirroring in the toxicity model the efficacy covariate terms in (4), yielding a model with nine parameters including ψ. The extra model complexity reduces approval probabilities and coverage. Poor coverage is a particular problem in the toxicity model in scenarios where the event rate is 10%. For instance, the four-parameter model performs very poorly in scenarios 1 and 3, particularly in the smallest cohorts. Performance is better in scenario 2 where the true rate is 30%. This is notable because the published data [9, 10] suggest low toxicity. In scenarios not shown in Table 3, this model successfully identifies differential toxicity associated with covariates but requires a

Sc	Co	PrEff	PrTox	OddsR	N	Eff	Tox	Inf	Reg	Diff	BetaBin
1	1	0.300	0.1	1.0	9.3	2.8	0.9	0.883	0.896	0.878	0.540
	2	0.300	0.1	1.0	13.1	3.9	1.3	0.906	0.920	0.905	0.658
	3	0.300	0.1	1.0	7.5	2.3	0.8	0.980	0.909	0.816	0.473
	4	0.300	0.1	1.0	12.5	3.7	1.2	0.875	0.912	0.896	0.635
	5	0.300	0.1	1.0	10.8	3.2	1.1	0.873	0.909	0.890	0.590
	6	0.300	0.1	1.0	6.8	2.0	0.7	0.959	0.893	0.819	0.459
2	1	0.100	0.3	1.0	9.3	0.9	2.8	0.012	0.025	0.019	0.035
	2	0.100	0.3	1.0	13.1	1.3	3.9	0.013	0.028	0.023	0.032
	3	0.100	0.3	1.0	7.5	0.8	2.3	0.038	0.029	0.021	0.034
	4	0.100	0.3	1.0	12.5	1.2	3.7	0.009	0.024	0.021	0.034
	5	0.100	0.3	1.0	10.8	1.1	3.2	0.009	0.024	0.022	0.032
	6	0.100	0.3	1.0	6.8	0.7	2.0	0.027	0.025	0.019	0.041
3	1	0.300	0.1	0.2	9.3	2.8	0.9	0.884	0.897	0.879	0.562
	2	0.300	0.1	0.2	13.1	3.9	1.3	0.906	0.920	0.904	0.667
	3	0.300	0.1	0.2	7.5	2.3	0.8	0.981	0.909	0.818	0.494
	4	0.300	0.1	0.2	12.5	3.7	1.2	0.877	0.913	0.897	0.652
	5	0.300	0.1	0.2	10.8	3.2	1.1	0.874	0.908	0.889	0.605
	6	0.300	0.1	0.2	6.8	2.0	0.7	0.960	0.893	0.820	0.478
4	1	0.167	0.1	1.0	9.3	1.5	0.9	0.408	0.451	0.398	0.293
	2	0.192	0.1	1.0	13.1	2.5	1.3	0.651	0.690	0.633	0.432
	3	0.500	0.1	1.0	7.5	3.8	0.8	0.993	0.981	0.974	0.622
	4	0.091	0.1	1.0	12.5	1.1	1.3	0.208	0.277	0.215	0.131
	5	0.156	0.1	1.0	10.8	1.7	1.1	0.405	0.493	0.419	0.298
	6	0.439	0.1	1.0	6.8	3.0	0.7	0.961	0.930	0.931	0.581
5	1	0.167	0.3	1.0	9.3	1.5	2.8	0.027	0.063	0.039	0.071
	2	0.192	0.3	1.0	13.1	2.5	3.9	0.046	0.099	0.066	0.084
	3	0.500	0.3	1.0	7.5	3.8	2.3	0.071	0.141	0.102	0.159
	4	0.091	0.3	1.0	12.5	1.1	3.7	0.014	0.037	0.021	0.028
	5	0.156	0.3	1.0	10.8	1.7	3.2	0.030	0.071	0.045	0.065
	6	0.439	0.3	1.0	6.8	3.0	2.0	0.070	0.135	0.099	0.163
6	1	0.167	0.1	0.2	9.3	1.5	0.9	0.408	0.451	0.396	0.308
	2	0.192	0.1	0.2	13.1	2.5	1.3	0.651	0.689	0.633	0.447
	3	0.500	0.1	0.2	7.5	3.8	0.8	0.993	0.981	0.974	0.627
	4	0.091	0.1	0.2	12.5	1.1	1.3	0.208	0.278	0.212	0.139
	5	0.156	0.1	0.2	10.8	1.7	1.1	0.402	0.493	0.415	0.313
	6	0.439	0.1	0.2	6.8	3.0	0.7	0.962	0.929	0.930	0.589

Table 3 Summary of simulated trials. Sc is scenario number; Co is cohort number. Patient cohorts are defined in Table 2. PrEff and PrTox are true probabilities of efficacy and toxicity. OddsR shows ratio of odds of efficacy in patients that experience toxicity to those that do not. OddsR=1 reflects no association; OddsR$<$1 implies efficacy is less likely when toxicity occurs. N shows mean number of patients; Eff and Tox the mean number of events. Inf is the approval probability under informative priors; Reg and Diff are the same under regularising and diffuse priors. BetaBin shows approval probability using cohort-specific beta-binomial models. 10,000 iterations used.

sample size exceeding 100 to do so with high probability. Weighing the extra demand in resource against the likely benefit, we prefer the simpler model.

Lastly, scenarios 3 and 6 show that model performance is seemingly unaffected by strong association in efficacy and toxicity events. We investigated a model variant that assumes independence by setting $\psi = 0$ in (2). Approval probability and precision were practically unchanged. This is understandable because ψ is absent

from (4) and therefore does not affect (3). ψ is useful, however, in conditional inference. For example, the predicted distribution of unknown efficacy conditioned on observed toxicity is shifted lower by ψ given negative association prevailing in the collected trial data, and vice-versa. Given its useful role with no performance penalty, we retain ψ.

6 Further Work and Availability of Materials

Statisticians know that dichotomising continuous variables reduces information. We have used in this research the PD-L1 categorisation previously introduced and validated in NSCLC [9] . In ongoing work, we use the underlying continuous score in place of the categorisation. In this setting, further care must be taken when specifying the model form and the parameter priors. For instance, we expect overwhelmingly that the gradient term describing the sensitivity of efficacy with respect to PD-L1 score will be positive, so that higher scores are more likely to yield efficacy events. However, it is debatable whether our priors or model form should reflect that we expect greater or lesser efficacy-PD-L1 sensitivity in treatment naive or pretreated patients. A hierarchical approach has some merit, where PD-L1 gradients are interpreted as draws from some common distribution. This would allow heterogeneity to manifest in subgroups whilst discouraging over-fitting via shrinkage-based regularisation. Missing data is a perennial challenge in clinical trials. A hierarchical approach has the further benefit of pragmatically treating patients with unknown pretreatment status as a third cohort. Intuitively, we could interpret this group as behaving like an unknown mixture of pretreated and treatment-naive patients.

One of the focuses of this research has been the consideration of different models that could eventually be fit to the trial data. We approached the problem as if one candidate model had to be identified in advance in the analysis plan. An alternative is to specify a suite of models and then combine their inferences. For instance, in Bayesian model averaging, the response distributions generated by the candidate models are weighted together by their marginal posterior probabilities. In contrast, methods have been introduced that *stack* posterior predictive distributions, using the leave-one-out (LOO) predictor for each model and each data-point, deriving model weights that minimise the LOO mean squared error [18] . A method like this could allow us to combine models with markedly different features like simple and complex specifications for the toxicity sub-model in a data-oriented manner.

Models used in this research were implemented in Stan [5] and all materials are available on GitHub at https://github.com/brockk/bebop

References

1. Borghaei, H., Paz-Ares, L., Horn, L., Spigel, D.R., Steins, M., Ready, N.E., et al.: Nivolumab versus docetaxel in advanced non-squamous non-small-cell lung cancer. N Engl J Med. **373** 123–135 (2015)
2. Braun, T.M.: The bivariate continual reassessment method: extending the CRM to phase I trials of two competing outcomes. Control Clin Trials. **23** 240–256 (2002)
3. Brutti, P., Gubbiotti, S., Sambucini, V.: An extension of the single threshold design for monitoring efficacy and safety in phase II clinical trials. Statist. Med. **30** 1648–1664 (2011)
4. Bryant, J., Day, R.: Incorporating toxicity considerations into the design of two-stage phase II clinical trials. Biometrics. **51** 1372–1383 (1995)
5. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., et al.: Stan: a probabilistic programming language. J. Stat. Softw. **76** 1–32 (2017)
6. Conaway, M., Petroni, G.: Designs for phase II trials allowing for a trade-off between response and toxicity. Biometrics. **52** 1375–1386 (1996)
7. Cook, R., Farewell, V.: Guidelines for monitoring efficacy and toxicity responses in clinical trials. Biometrics. **50** 1146–1152 (1994)
8. Eisenhauer, E.A., Therasse, P., Bogaerts, J., Schwartz, L.H., Sargent, D., et al.: New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur. J. Cancer. **45** 228–247 (2009)
9. Garon, E.B., Rizvi, N.A., Hui, R., Leighl, N., Balmanoukian, A.S., et al.: Pembrolizumab for the treatment of nonsmall-cell lung cancer. N Engl J Med. **372** 2018–2028 (2015)
10. Herbst, R.S., Baas, P., Kim, D., Felip, E., Perez-Gracia,J.L., et al.: Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. Lancet. **387** 1540–1550 (2016)
11. Jin, H.: Alternative designs of phase II trials considering response and toxicity. Contemp Clin Trials. **109** 525–536 (2007)
12. Konopleva, M., Thall, P.F., Arana Yi, C., Borthakur, G., Coveler, A., et al.: Phase I/II study of the hypoxia-activated prodrug PR104 in refractory/relapsed acute myeloid leukemia and acute lymphoblastic leukemia. Haematologica. **100** 927–934 (2015)
13. Schiller J.H., Harrington D., Belani C.P., Langer C., Sandler A., et al.: Comparison of four chemotherapy regimens for advanced non-small-cell lung cancer. N Engl J Med. **346** 92–98 (2002)
14. Thall, P. F., Simon, R. M., Estey, E. H.: New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. J. Clin. Oncol. **14** 296–303 (1996)
15. Thall, P. F., Sung, H. G.: Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. Statist. Med. **17** 1563–1580 (1998)
16. Thall, P.F., Cook, J.D.: Dose-finding based on efficacy-toxicity trade-offs. Biometrics. **60** 684–693 (2004)
17. Thall, P.F., Nguyen, H.Q., Estey, E.: Patient-specific dose finding based on bivariate outcomes and covariates. Biometrics. **64** 1126–1136 (2008)
18. Yao, Y., Vehtari, A., Simpson, D., Gelman, A.: Using stacking to average Bayesian predictive distributions. Bayesian Anal. **13** 917–1007 (2017)
19. Zhang, W., Sargent, D.J., Mandrekar, S.: An adaptive dose-finding design incorporating both toxicity and efficacy. Statist. Med. **25** 2365–2383 (2006)